Advanced data visualizations with Stata

#StataViz ++

Asjad Naqvi

International Institute for Applied Systems Analysis (IIASA)
Wirtschaftsuniversität Wien (WU)

Stata UK Conference
10 September, 2021
Background

- Ph.D. Economics (2007-12), New School for Social Research, New York, USA.
- Started using Stata during my M.Sc. Economics degree (2003-04)
- Worked/still working on a ton of research projects mostly using Stata
  - Overseeing several large data projects
  - Going through more code than I like
- Why do all the dataviz stuff?
  - Online COVID-19 dataviz explosion + work-from-home → curiosity
  - Huge positive response from the online community
Why we need more #StataViz?

- Few official releases for graphs
- Still lots of options available in the default Stata structure
- Lots of great development in terms of dataviz packages:
  - colorpalette, colrspace, heatplot (Ben Jann), spmap (Maurizio Pisati), various packages by Nick Cox, nwcommands (Thomas Grund)
Why we need more #StataViz?

- Few official releases for graphs
- Still lots of options available in the default Stata structure
- Lots of great development in terms of dataviz packages:
  - colorpalette, colrspace, heatplot (Ben Jann), spmap (Maurizio Pisati), various packages by Nick Cox, nwcommands (Thomas Grund)
- What I will discuss here:
  - What has been developed
  - What is still possible
  - What core elements Stata should still add
COVID-19 cumulative deaths per million population (1 Mar 21)

Data source: Our World in Data COVID-19 tracker. Total deaths for the last reported date given in brackets.

Guide: Arrows
colorpalette + stacked-area graphs

New cases - European countries

Total for today = 72,670

Daily cases (7-days moving average)

Data sources: Our World in Data, JHU, ECDC. World Bank classifications used for country groups. Top 10 countries are labeled.

Guide: Stacked area graphs
colorpalette + stacked-area graphs

Share of new cases - European countries

Total for today = 72,670

Data sources: Our World in Data. World Bank classifications used for country groups. Top 10 countries are labeled.

Guide: Stacked area graphs
COVID-19 daily global deaths: 9,130

Guide: Stream plots

Data sources: Our World in Data, World Bank 2020 classifications used for country groups.
Ridgeline (Joy) plots

COVID-19 daily deaths in Europe

Data sources: Our World in Data. World Bank classifications used for country groups. Each country plot is normalized by its maximum value.

Guide: Joy plots
**COVID-19 Policy Stringency Index**


Guide: Heat plots
COVID-19 Policy Stringency Index

Guide: Heat plots

Scatter plot
Scheme neon

The Stata Guide
The motivation
Colors, Lines, and Areas
Maps
Polar
Unconventional
In the pipeline
Last slides

Schemes + colorpalette

TSG Schemes
The motivation

Colors, Lines, and Areas

Maps

Polar

Unconventional

In the pipeline

Last slides

spmap + colorpalette

Guide: Maps II

Median age in Europe by NUTS 3 regions

Dependency Ratio in Europe by NUTS 3 regions

Data source: Eurostat table: demo_j_ganm10, NUTS 2016 layers from Eurostat (GSDC).

Data source: Eurostat table: demo_j_ganm10, NUTS 2016 layers from Eurostat (GSDC).
OSM + QGIS to Stata

- GIS data from OpenStreetMaps (OSM)
- **QGIS** can be used to extract multiple layers (points, lines, polygons)
- Why do this?
  - Fully customized maps
  - Spatial layer clipping + control over bounding box
  - Allow for very rich visualizations
Vienna

Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS
Vienna

Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS
New York

Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS
Maps + Networks

Social Connectivity Index - Austria (70+ pctl)

Social Connectivity Index - UK (70+ pctl)

Map layers: GISCO Eurostat, SCI: Facebook. Within country ties not included. Darker shade is higher SCI.
The motivation
Colors, Lines, and Areas
Maps
Polar
Unconventional
In the pipeline
Last slides

$x = r \cos \theta$
$y = r \sin \theta$
$r^2 = x^2 + y^2$
$\theta = \text{atan2}(y, x)$

180 degrees = $\pi$ radians
Spider plots

COVID-19 policies: 29 Aug 2021


Guide: Stata graphs: Spider plots
Polar plots

Excess weekly deaths for 65+ - CZ


Guide: Stata graphs: Polar plots
Coxcomb plots (by Florence Nightingale)

Guide: Stata graphs: Rose (Coxcomb) plots
The Roving Mind (1983)

Isaac Asimov, The Roving Mind (1983)

If there is evidence for it, the wilder and more ridiculous, confirmed by independent observers, I believe in observational measurement. And ridiculous reasoning, I believe in evidence. However, no matter how wild and more solid, the evidence will have to be.
Hypotrochoids (spiralographs), Guilloché patterns
The Du Bois Challenge

Original

Replication

The Du Bois Challenge

Original

Replication

The Du Bois Challenge

Relational and hierarchical datasets

- Both are in a network structure
- *Relational*: flows from one node to another for all layers
- *Hierarchical*: Flows are partitioned and ordered by layers
<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>App</td>
<td>Blog</td>
<td>325</td>
</tr>
<tr>
<td>App</td>
<td>Homepage</td>
<td>255</td>
</tr>
<tr>
<td>App</td>
<td>Medium</td>
<td>220</td>
</tr>
<tr>
<td>App</td>
<td>Twitter</td>
<td>180</td>
</tr>
<tr>
<td>Blog</td>
<td>App</td>
<td>144</td>
</tr>
<tr>
<td>Blog</td>
<td>Homepage</td>
<td>184</td>
</tr>
<tr>
<td>Blog</td>
<td>Twitter</td>
<td>320</td>
</tr>
<tr>
<td>Blog</td>
<td>Website</td>
<td>50</td>
</tr>
<tr>
<td>Direct</td>
<td>App</td>
<td>124</td>
</tr>
<tr>
<td>Direct</td>
<td>Blog</td>
<td>175</td>
</tr>
<tr>
<td>Direct</td>
<td>Homepage</td>
<td>119</td>
</tr>
<tr>
<td>Facebook</td>
<td>App</td>
<td>124</td>
</tr>
<tr>
<td>Facebook</td>
<td>Blog</td>
<td>167</td>
</tr>
<tr>
<td>Facebook</td>
<td>Homepage</td>
<td>185</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google search</td>
<td>App</td>
<td>233</td>
</tr>
<tr>
<td>Google search</td>
<td>Blog</td>
<td>252</td>
</tr>
<tr>
<td>Google search</td>
<td>Homepage</td>
<td>131</td>
</tr>
<tr>
<td>Homepage</td>
<td>App</td>
<td>375</td>
</tr>
<tr>
<td>Homepage</td>
<td>Blog</td>
<td>287</td>
</tr>
<tr>
<td>Homepage</td>
<td>Twitter</td>
<td>50</td>
</tr>
<tr>
<td>Homepage</td>
<td>Website</td>
<td>100</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>App</td>
<td>185</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>Blog</td>
<td>173</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>Homepage</td>
<td>123</td>
</tr>
<tr>
<td>Twitter</td>
<td>App</td>
<td>485</td>
</tr>
<tr>
<td>Twitter</td>
<td>Blog</td>
<td>341</td>
</tr>
<tr>
<td>Twitter</td>
<td>Homepage</td>
<td>219</td>
</tr>
</tbody>
</table>
A matrix representation

Package: heatplot
A network representation (circle layout)

Package: nwcommands
A network representation (MDS layout)

Package: nwcommands. MDS = Multi-Dimensional Scaling
### Sample data - Hierarchical

<table>
<thead>
<tr>
<th>Layer</th>
<th>From</th>
<th>To</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Twitter</td>
<td>Blog</td>
<td>341</td>
</tr>
<tr>
<td>1</td>
<td>Twitter</td>
<td>Homepage</td>
<td>219</td>
</tr>
<tr>
<td>1</td>
<td>Twitter</td>
<td>App</td>
<td>485</td>
</tr>
<tr>
<td>1</td>
<td>Google search</td>
<td>Blog</td>
<td>252</td>
</tr>
<tr>
<td>1</td>
<td>Google search</td>
<td>Homepage</td>
<td>131</td>
</tr>
<tr>
<td>1</td>
<td>Google search</td>
<td>App</td>
<td>233</td>
</tr>
<tr>
<td>1</td>
<td>Facebook</td>
<td>Blog</td>
<td>167</td>
</tr>
<tr>
<td>1</td>
<td>Facebook</td>
<td>Homepage</td>
<td>185</td>
</tr>
<tr>
<td>1</td>
<td>Facebook</td>
<td>App</td>
<td>124</td>
</tr>
<tr>
<td>1</td>
<td>LinkedIn</td>
<td>Blog</td>
<td>173</td>
</tr>
<tr>
<td>1</td>
<td>LinkedIn</td>
<td>Homepage</td>
<td>123</td>
</tr>
<tr>
<td>1</td>
<td>LinkedIn</td>
<td>App</td>
<td>185</td>
</tr>
<tr>
<td>1</td>
<td>Direct</td>
<td>Blog</td>
<td>175</td>
</tr>
<tr>
<td>1</td>
<td>Direct</td>
<td>Homepage</td>
<td>119</td>
</tr>
<tr>
<td>1</td>
<td>Direct</td>
<td>App</td>
<td>124</td>
</tr>
<tr>
<td>2</td>
<td>Blog</td>
<td>Homepage</td>
<td>184</td>
</tr>
<tr>
<td>2</td>
<td>Blog</td>
<td>App</td>
<td>144</td>
</tr>
<tr>
<td>2</td>
<td>Homepage</td>
<td>Blog</td>
<td>287</td>
</tr>
<tr>
<td>2</td>
<td>Homepage</td>
<td>App</td>
<td>375</td>
</tr>
<tr>
<td>2</td>
<td>App</td>
<td>Homepage</td>
<td>255</td>
</tr>
<tr>
<td>2</td>
<td>App</td>
<td>Blog</td>
<td>325</td>
</tr>
<tr>
<td>3</td>
<td>Homepage</td>
<td>Website</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>Homepage</td>
<td>Twitter</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>App</td>
<td>Twitter</td>
<td>180</td>
</tr>
<tr>
<td>3</td>
<td>App</td>
<td>Medium</td>
<td>220</td>
</tr>
<tr>
<td>3</td>
<td>Blog</td>
<td>Twitter</td>
<td>320</td>
</tr>
<tr>
<td>3</td>
<td>Blog</td>
<td>Website</td>
<td>50</td>
</tr>
</tbody>
</table>
A matrix representation

Package: heatplot
Arc
Arc
Sankey
Why do all of this in Stata?

- One can pass the DataViz stuff to more established languages:
  - R using `rcall` (Haghish 2019) or Python using `PyStata`
  - But these should be viewed more as stop-gap options
  - Develop and internalize the dataviz option within Stata
  - Why?
Why do all of this in Stata?

- One can pass the DataViz stuff to more established languages:
  - R using `rcall` (Haghish 2019) or Python using `PyStata`
  - But these should be viewed more as stop-gap options
  - Develop and internalize the dataviz option within Stata
  - Why?

- Easy to integrate with the Stata code workflow
- Reduce the language burden on end-users
- Time investment? One-time (high) sunk costs or
- Some of the #StataViz shown above can be easily packaged for end-users
Core features that can take Stata to the next level (wish list)

- Ability to control marker size scaling
- Line weights (like marker weights)
- Allow angles/sizes/colors to be read from variables (like mlabels)
- Color/alpha scaling for markers/lines
- Color gradients for lines/areas
- Ability to add custom markers (or increase marker pool)
- Ability to added colored text in graphs
- Ability to add images in figures
- Ability to read images (pixel data/colors)
- Ability to draw/add/subtract areas between two or more functions
- Line curvature
Thank you!

For more #Stataviz:

- The Stata Guide on Medium
- GitHub: asjadnaqvi
- Twitter: @AsjadNaqvi

A lot more is in development for next year